DATA PIPELINES AND WORKFLOWS FOR CLOUD AND HYBRID ENVIRONMENTS

Infoworks

The Challenge

Data operations, orchestration and analytics are moving to the cloud. Cheap storage and computational costs are pay-as you-go, making initial startup costs more affordable than on-premise. Plus, big data and data warehouse environments in the cloud are relatively easier to use compared to on-premise offerings. However, there are still considerable challenges that you have to overcome to successfully implement and manage enterprise-class data operations at scale. And if you are moving to the cloud, you need to first take into account how you will:

- Migrate and synchronize large volumes of data
- Migrate existing data pipelines and quickly develop new pipelines
- Create enterprise-class, production-ready batch, incremental and streaming data pipelines that run at scale
- Manage your data operations across multi-cloud and hybrid deployments
- Progress your data operations to keep up with the never-ending evolution of underlying compute and storage environments

The Solution

Infoworks DataFoundry and DataReplicator automate data migration, as well as the development, management and orchestration of data pipelines and workflows for delivering data & analytics projects at scale, both on-premise and in the cloud.

Infoworks DataReplicator provides high-performance replication of data and metadata for large data clusters for on-premise to on-premise, on-premise to cloud, and cloud to cloud scenarios. It is an automated, production-ready platform for bi-directional, fault tolerant data replication that removes the need for custom scripting.

Infoworks DataFoundry automates the development and operationalization of data pipelines from source to consumption in support of business intelligence (BI), machine learning (ML) and artificial intelligence (AI) analytics applications. Infoworks DataFoundry integrates data ingestion, pipeline design, data access management, monitoring and orchestration into a fully unified system for delivering data, data pipelines and data workflows at scale, on any big data platform, in the cloud or on-premise.

CASE STUDY:
FINANCIAL SERVICES FIRM

Objective

- Migrate existing data warehouse data, pipelines and workflows into a new big data environment

Alternatives

- The company evaluated open source tooling that required many months of services, was limited in functionality and could not do bi-directional synch of data or automated conversion of existing SQL pipelines

Initial Results

Initial project migrated production scale workloads with 1.8 trillion records process, 80 complex pipelines and 2 complex workflows

- Migrated 90% of SQL pipeline logic automatically
- Migrated legacy Netezza workloads using drag and drop orchestration manager
- Migrated all initial data and workloads in 3 weeks with 1 full time resource with 99.999% accuracy

info@infoworks.io

(650) 391-9306

Infoworks

Capabilities Summary

CAPABILITY	DESCRIPTION
Incremental, bidirectional data migration and replication	Migrate data from on-premise clusters to the cloud, or cloud to cloud, and keep multiple data clusters synchronized. Provides high-performance replication of data and metadata for large data volumes and a variety of scenarios
Batch, CDC and Streaming Data Ingestion	Ingest source data in a high-performance, parallel process, while automatically preserving data precision. DataFoundry provides a no-code environment for configuring the ingestion of data into your data lake via batch, change data capture (CDC), and data streaming.
Data pipeline development and migration	Provides self-service data preparation using an interactive, drag-and-drop data transformation capability with support for automated migration of pre-existing SQL-based transformation logic. Users work with data in a collaborative, suggestion-based interface that reduces or eliminates dependence on IT skills. DataFoundry makes it possible to integrate data pipelines with advanced analytics algorithms from libraries such as SparkML and R, without the need for coding. Build trained models or import pre-trained models into data pipelines.
Data Workflow Operationalization	Design end-to-end workflows and orchestrate them in production with fault-tolerant, distributed execution. Migrate from development environments to production across big data or cloud platforms with single-click operation.
Cross-cloud and onpremise orchestration & management	Design end-to-end workflows and orchestrate them in production with fault-tolerant, distributed execution. Push development pipelines and workflows to production across big data or cloud platforms with single-click operation.
Data workflow portability	Data workflows developed in DataFoundry can be run on a wide variety of execution engines and storage platforms. Workflows can be migrated from an on-premise Hadoop or Spark platform to the cloud, or from one cloud environment to another without recoding. DataFoundry automatically optimizes data workflows to run at scale on all supported execution engines.

About Infoworks

Infoworks provides the first Enterprise Data Operations and Orchestration (EDO2) software system to automate the development and operationalization of data pipelines from source to consumption in support of business intelligence (BI), machine learning (ML) and artificial intelligence (AI) applications. Infoworks' code-free development environment allows organizations to develop and manage end-to-end data workflows, or migrate existing data and workflows, without requiring an army of big data experts. Infoworks delivers capabilities to automate and simplify development of data ingestion, data preparation, query acceleration and ongoing operationalization of production data pipelines at scale. Infoworks supports cloud, multi-cloud, and on premise deployments, enabling customers to deploy projects to production within days, dramatically increasing business agility and accelerating time to value.